HPGP Analysis

Margaret Janiczek

2024-04-21

Data exploration

Originally there were 629 subjects, 5000 SNPs, and several features such as astigmatism, high triglycerides, gender weight and height.

Characteristic	**N = 629**
High_triglycerides	46 (11%)
Unknown	219
Asthma	81 (21%)
Unknown	245
Irritable_bowel_syndrome	65 (16%)
Unknown	234
Non_melanoma_skin_cancer	27 (6.5%)
Unknown	213
Astigmatism	158 (40%)
Unknown	231
Iron_deficiency_anemia	79 (20%)
Unknown	225
Myopia	221 (56%)
Unknown	231
Ovarian_cysts	39 (10%)
Unknown	246
Presbyopia	50 (13%)
Unknown	231
Osteoarthritis	51 (13%)
Unknown	243
High_cholesterol	90 (22%)
Unknown	219
Hypertension	64 (16%)
Unknown	237
Colon_polyps	39 (9.4%)
Unknown	212
Gender	
Female	184 (35%)
Male	349 (65%)
Unknown	96
Weight	75 (64, 88)
Unknown	388
Height	175 (170, 180)
Unknown	403

As we can see there are 231 subjects missing values for astygmatism. I am going to exclude them from our analysis dataset. Additionally, there were 10 subjects that were missing >75% of SNP data so they will be excluded as well. This results in a sample size of 388 subjects.

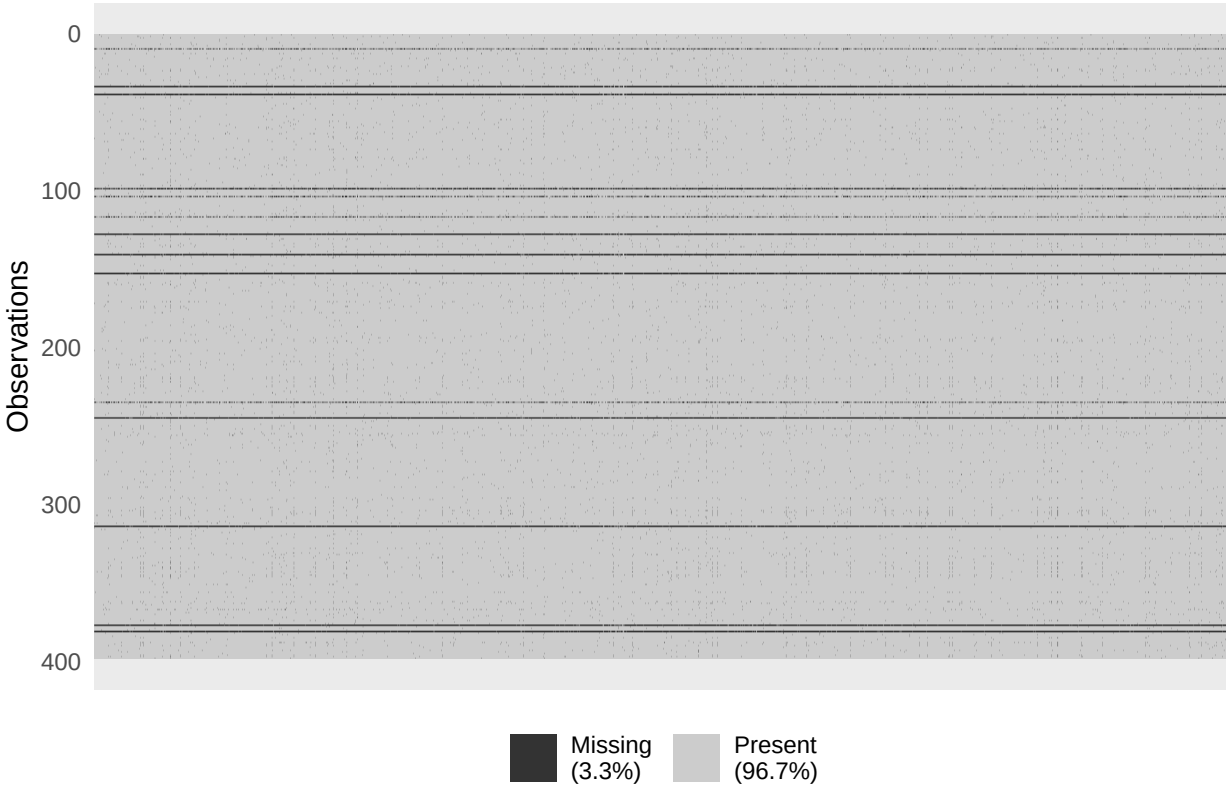
In the below table we can see the distribution of gender and myopia stratified by presence of astigmatism (1) vs no astigmatism (0).

Characteristic	**Overall**, N = 388	**0**, N = 233	**1**, N = 155	**p-value**
Myopia	215 (55%)	100 (43%)	115 (74%)	<0.001
Gender				0.13
Female	140 (37%)	77 (34%)	63 (41%)	
Male	240 (63%)	151 (66%)	89 (59%)	
Unknown	8	5	3	
Weight	75 (64, 87)	74 (64, 85)	76 (64, 91)	0.5
Unknown	231	137	94	
Height	175 (167, 180)	175 (167, 180)	175 (167, 181)	0.5
Unknown	242	142	100	

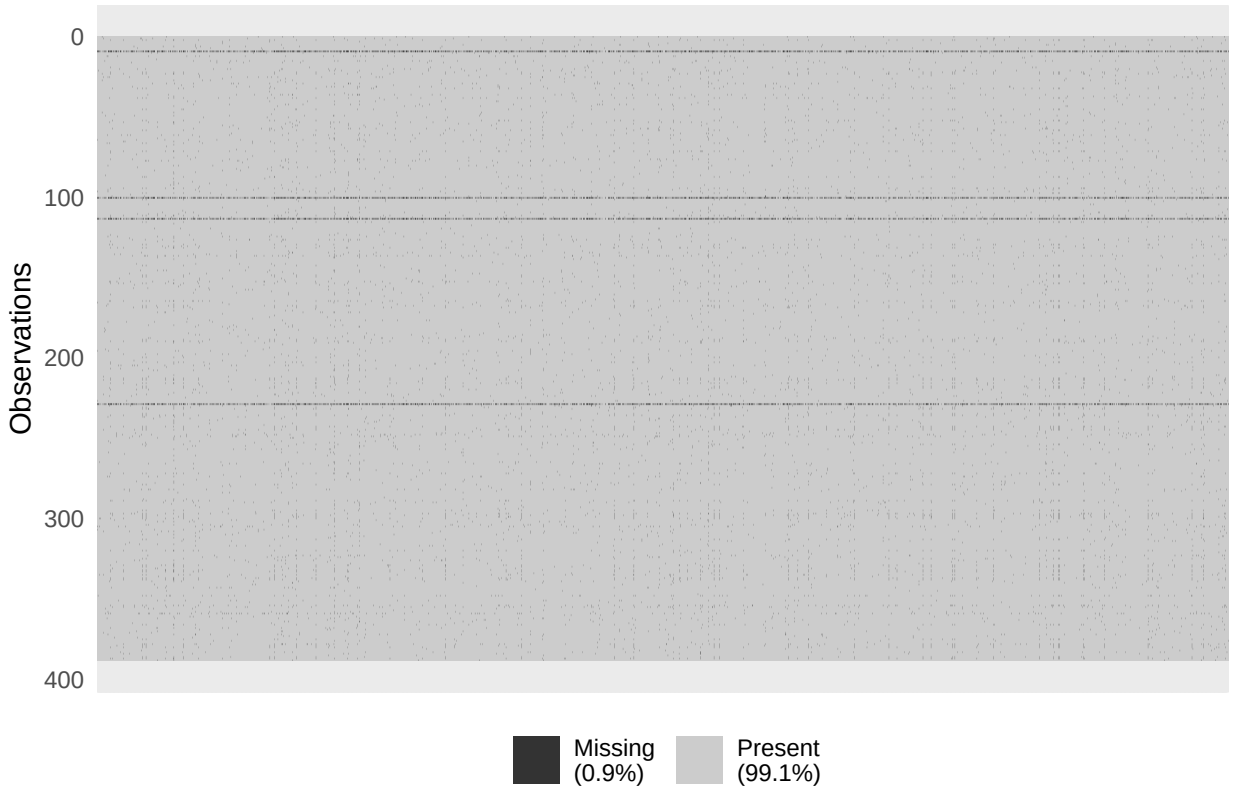
Next I calculated the percent of SNPs which had missingness. No SNPs were missing >75% of values, so we didn't need to filter out any SNPs for missingness. Finally, calculated the minor allele frequency (MAF) for all SNPs and excluded SNPs with MAF <5% for a final count of 4789 SNPs.

Using this "final" (pending agreement on filtering) data, see below for the distribution of the remaining missingness.

Firs plot: distribution of missingness prior to processing the original 398 subjects and 5000 SNPs:

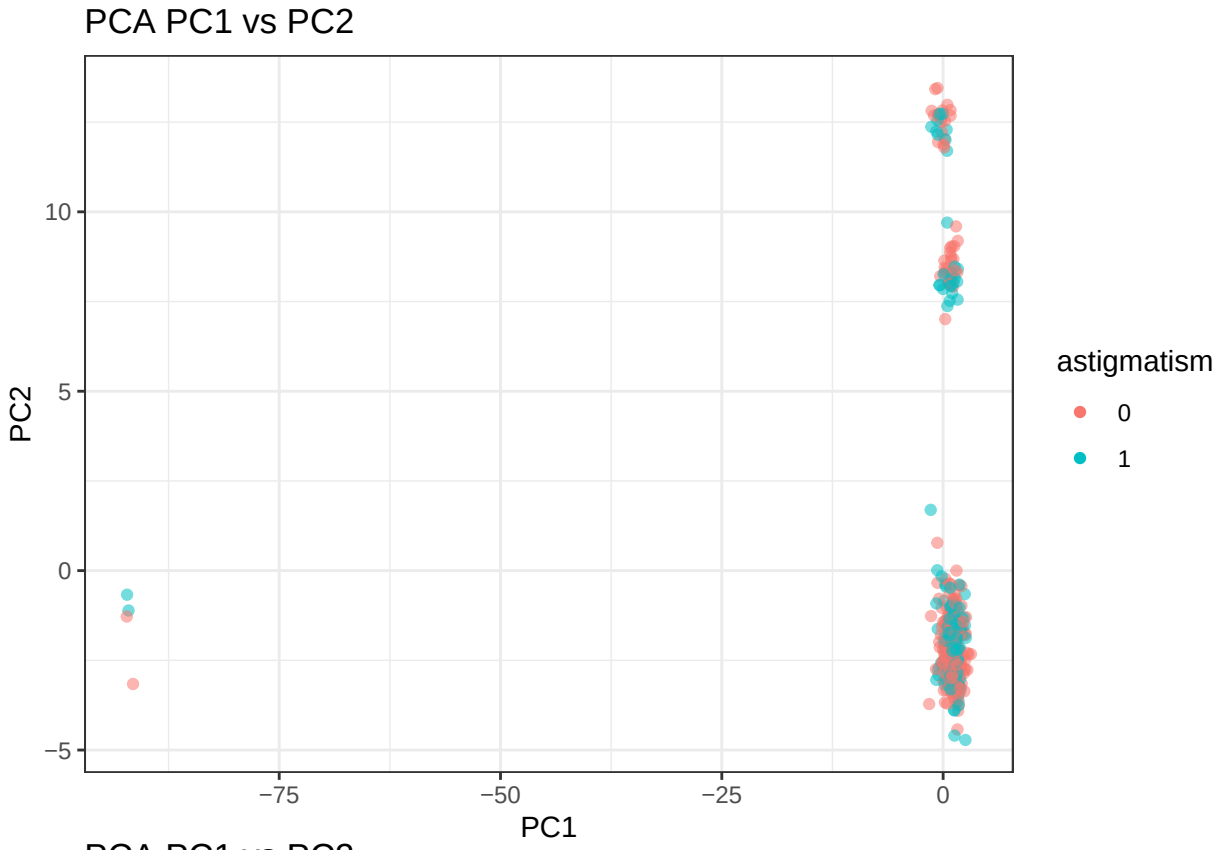


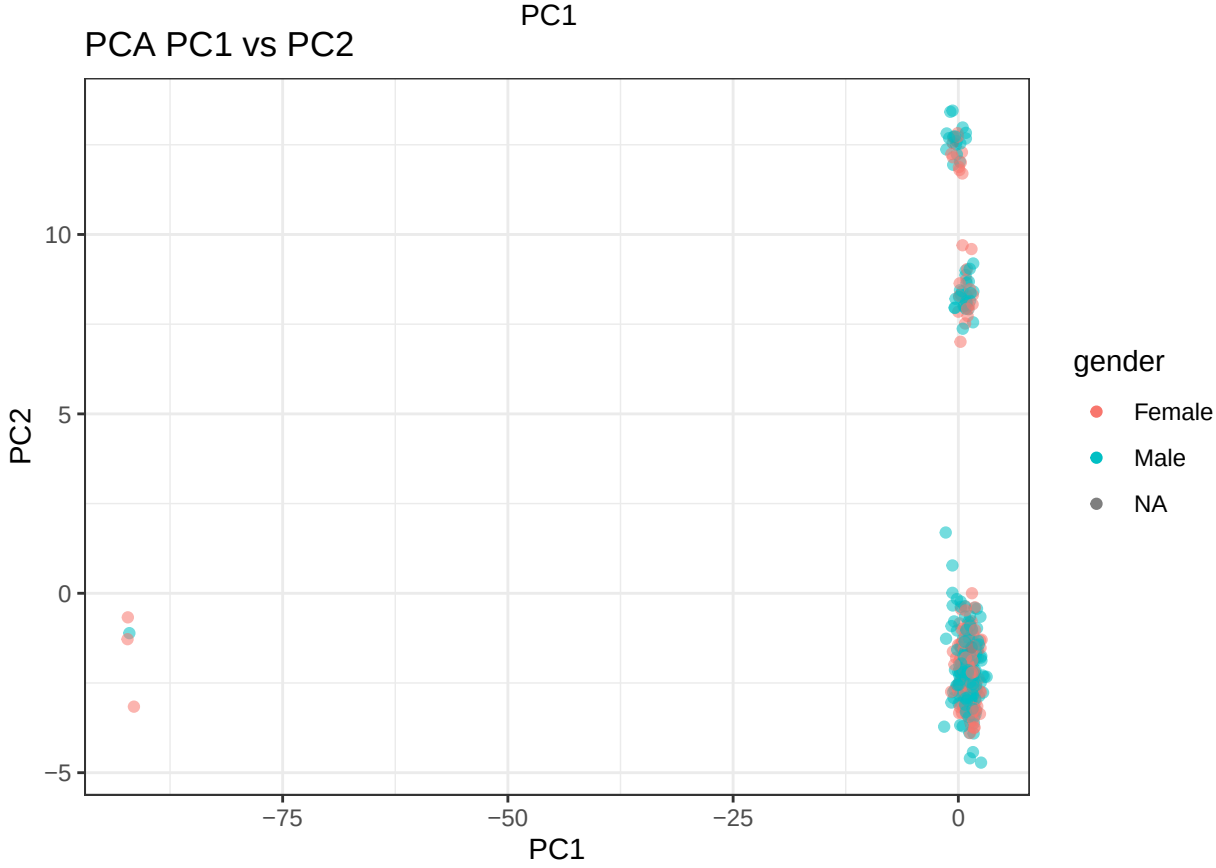
And below is the distribution of remaining missingness in 388 subjects for 4789 SNPs:



PCA

There appears to be some population substructure. However I'm not sure what it is related to in this dataset, since most of the variables are binary but this has 3 distinct groups.


PCA PC1 vs PC2
PC2
PC1
10
5
0
−5
−75
−50
−25
0
astigmatism
0
1


PCA PC1 vs PC2
PC2
PC1
10
5
0
−5
−75
−50
−25
0
gender
Female
Male
NA